PSET A

Kate Little

2024-09-04

Load Packages

```
library(sjPlot)     # for tab_model()
library(vtable)     # for sumtable()
```

```
## Loading required package: kableExtra
```

```
library(tidyverse)  # for general data processing
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter()     masks stats::filter()
## ✖ dplyr::group_rows() masks kableExtra::group_rows()
## ✖ dplyr::lag()        masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
setwd("C:/Users/littkate/Desktop/EDUC7456/psetA")
```

Data

In this problem set we will use a subset of data from the Stanford Education Data Archive (SEDA) (https://edopportunity.org), part of the Educational Opportunity Project. Please remember to complete the data use agreement here (https://edopportunity.org/get-the-data/), especially if you are interested in using the SEDA data in other projects. The SEDA data includes data on educational opportunities in the US for nearly all public schools and public school districts enrolling students in grades 3 – 8. The complete data files include information on student achievement test scores from state accountability testing as well as socio-demographic information from sources such as the NCES Common Core of Data (https://nces.ed.gov/ccd/) and the Census Bureau's American Community Survey (https://www.census.gov/programs-surveys/acs). These data represent average characteristics for each district during the time period 2009-2018.

In this problem set we will use district-level data. The website contains data at other levels of aggregation, with additional variables. In the PSET A dataset, each row represents a single school district. Although the file contains extra variables, the variables we will use in the current assignment are:

- **sedalea**: unique school district ID variable
- **sedaleaname**: name of the school district district
- **stateabb**: state abbreviation for state where district is located
- **rural**: percent of students in the district enrolled in schools in rural locales
- **avgrdall**: average per-grade enrollment in the district
- **sesavgall**: average socioeconomic status (SES) of families living in the school district, based on a combination of six factors. This variable is standardized relative to the national distribution of SES across districts.
- **avg_math**: average mathematics test scores in the district among 3rd-8th graders, on the "GCS" scale. On this scale, a value of "3" indicates average scores at the 3rd grade level, a value of "4" at the 4th grade level, etc.
- **avg_rla**: average reading/language arts test scores in the district among 3rd-8th graders, on the "GCS" scale. On this scale, a value of "3" indicates average scores at the 3rd grade level, a value of "4" at the 4th grade level, etc.

(Note that I have re-named some of the variables for simplicity. These will not necessarily match the variable names in the SEDA files on the website. You can access the data cleaning code that converts from the raw SEDA files to the PSET data on Canvas.)

Question 1

Run the following code to load the data, subset to key variables and remove districts with missing data. No answers needed.

```
seda <- read.csv(file = "seda_psetA.csv")
# the select(.data=X, ...VARS... ) function selects the variables VARS from X
seda <- select(.data=seda,
               sedalea, sedaleaname, stateabb,
               avg_math, avg_rla,
               sesavgall, rural, avgrdall)
```

Here the `sumtable()` function is used to produce a helpful quick summary of the data.

```
sumtable(seda)
```

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
sedalea	12835	2982734	1474356	100002	1806750	4031425	7200030
avg_math	12591	5.6	1.1	-3.6	4.8	6.3	10
avg_rla	12589	5.6	1.1	-0.59	4.9	6.2	9.8
sesavgall	12346	0.33	0.85	-4.4	-0.16	0.88	2.9
rural	12793	0.56	0.44	0	0.049	1	1
avgrdall	12793	293	1189	1.1	34	239	73295

The following code will drop any rows with missing values in them. This is generally not a good way to handle missing data, but can be a useful tool to know about. We'll use it here for simplicity.

```
nrow(seda)
```

```
## [1] 12835
```

```
seda <- drop_na(seda)
nrow(seda)
```

```
## [1] 12257
```

Question 2

Create a table of summary statistics for the variables avg_math, avg_rla, rural, sesavgall. The table should include the following statistics for each variable: number of observations, mean, standard deviation, minimum, and maximum values.

```
sumtable(seda, vars = c('avg_math', 'avg_rla', 'rural', 'sesavgall'))
```

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
avg_math	12257	5.6	1.1	0.34	4.9	6.3	10
avg_rla	12257	5.6	1.1	-0.24	4.9	6.2	9.8
rural	12257	0.55	0.44	0	0.045	1	1
sesavgall	12257	0.33	0.85	-4.4	-0.16	0.88	2.9

Question 3

How many total school districts are represented in the data now that we have removed districts with missing data?

Your answer:

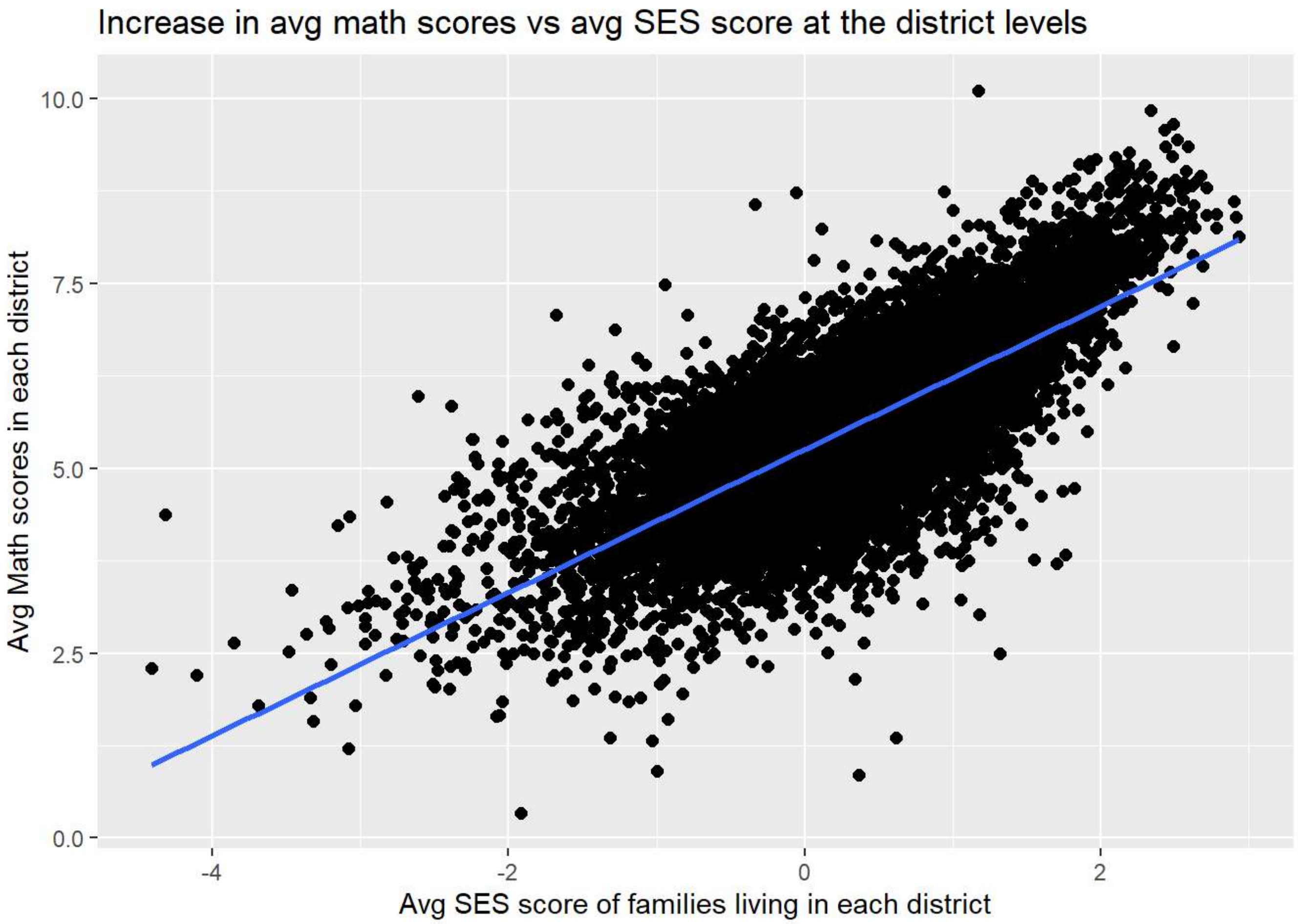
There are 12,257 districts with non-missing data in the dataset.

Question 4

Create a scatterplot showing average district SES on the X-axis and average math test scores on the Y-axis. Add a linear trend line to the scatterplot and calculate the correlation between average district SES and average math test scores. Based on the scatterplot and correlation briefly describe the association between these two variables.

```
seda %>%
  ggplot(aes(x=sesavgall, y=avg_math)) +
  geom_point(size=2) +
  geom_smooth(method="lm", se=FALSE)+
  ylab("Avg Math scores in each district")+
  xlab("Avg SES score of families living in each district")+
  ggtitle("Increase in avg math scores vs avg SES score at the district levels")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
cor(seda$sesavgall, seda$avg_math)
```

```
## [1] 0.7382288
```

Based on the scatterplot and the data there appears to be a strong positive correlation that appears to be linear between average SES score and average math score at the school district level. The correlation between avg SES score and avg math score is 0.73 indicating that as average SES scores increase, so do average math scores.

Question 5

Because these data represent the AVERAGE math test scores and AVERAGE family SES across districts, the correlation in #4 is an "ecological correlation." If we had INDIVIDUAL student-level test scores and family SES values, do you think the correlation between INDIVIDUAL student math test scores and family SES would be HIGHER, LOWER, or ABOUT THE SAME as the correlation in #4? Briefly explain your answer.

I would intuitively think that the correlation would be stronger, because of other determinants in household and behaviors. Those homes with higher socioeceonomic status would likely have much more facility to promote a child's math learning than those with lower SES. Although, the response to the previous question provides no statistical insight into trends at the individual level from the district aggregate level as ecological fallacy does not permit inferences from one ecological (aggregate) level to another level such as that of the individual.

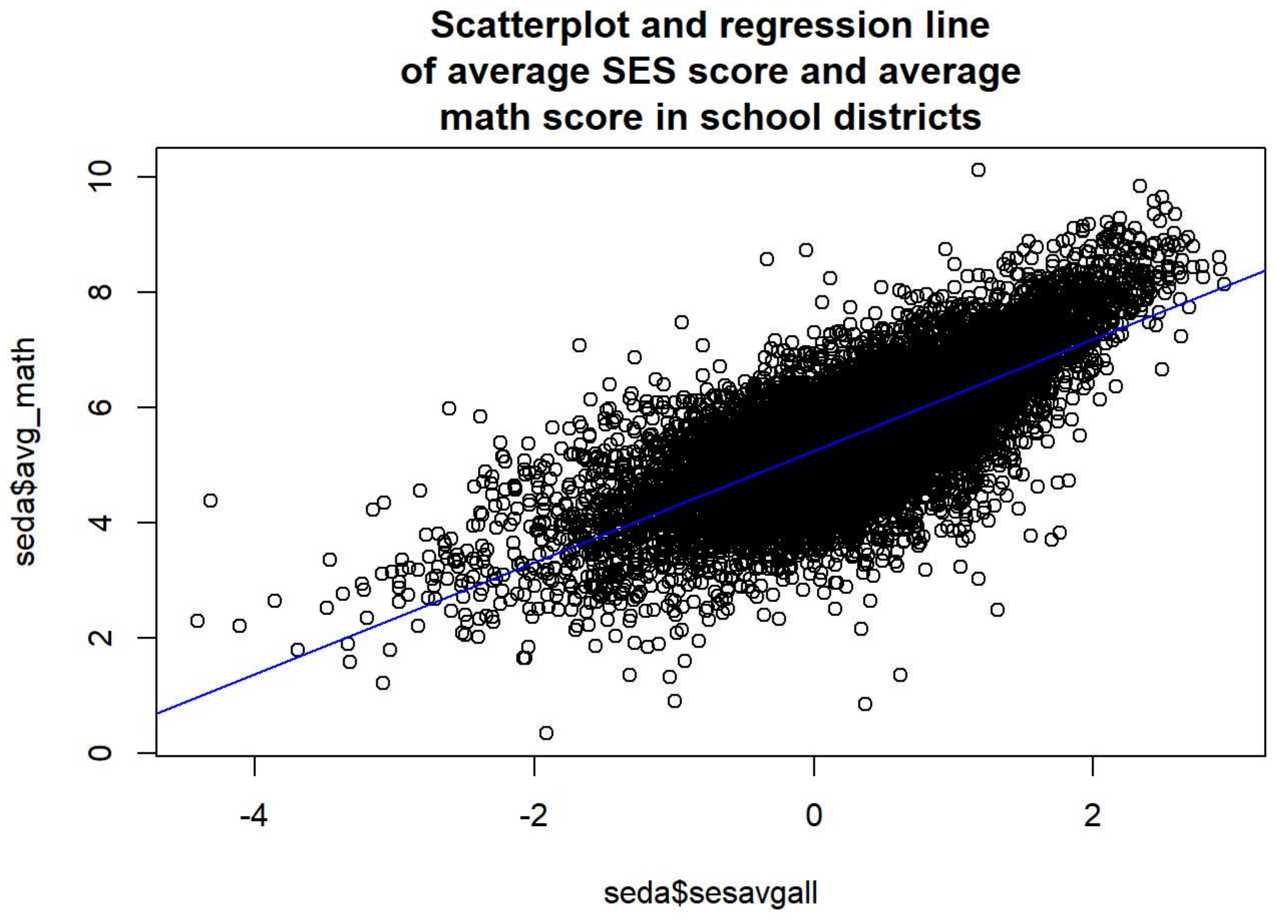
```
# no  code here
```

Question 6

Estimate a simple linear regression model using least squares that predicts district average math test scores as a function of district average SES. Produce summary information for the model using the built-in summary() command or a nice table with a function such as tab_model(). Provide a 1-sentence interpretation for each of the following statistics from the model: intercept, slope, r-squared, and root mean squared error (RMSE; R calls this the "residual standard error").

```
ses_math_model1<-lm(avg_math ~ sesavgall, data = seda)

plot(seda$sesavgall, seda$avg_math,main = "Scatterplot and regression line\nof average SES score and average\nmath score in school districts")
abline(ses_math_model1, col = "blue")
```



```
model_summary<-summary(ses_math_model1)
rmse<-model_summary$sigma
tab_model(ses_math_model1)
```

	avg math		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	5.25	5.24 – 5.27	**<0.001**

sesavgall	0.97	0.95 – 0.98	**<0.001**
Observations	12257		
R^2 / R^2 adjusted	0.545 / 0.545		

```
rmse
```

```
## [1] 0.7489835
```

```
# summary(ses_math_model1)
```

Interpretation

- For a district in which the average SES score is 0, the predicted average math GCS score would be 5.25.
- For a 1 unit change in SES score, the average math score is predicted to change by 0.97 units.
- 54.5% of the variation in increase in math scores is explained by the district’s SES score.
- The RMSE is 0.749 meaning that 95% of the actual values will fall within 2(0.749) SES units of the predicted value.

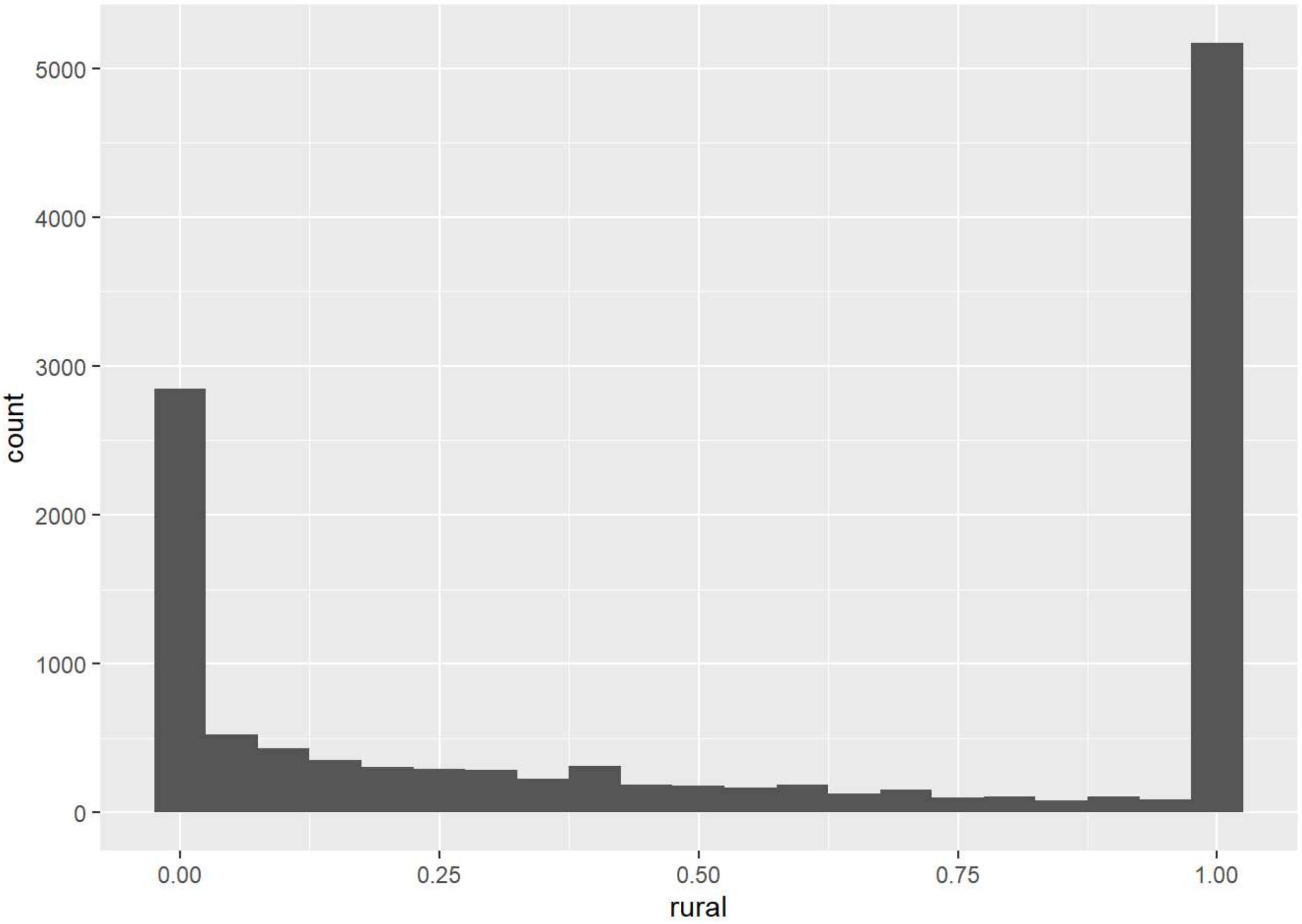
Question 7

Create a histogram of the “rural” variable. This variable reports the proportion of students in each district who are enrolled in rural schools. In 1-2 sentences briefly describe the shape of the distribution and explain why it might have this shape.

```
summary(seda$rural)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00000 0.04496 0.57371 0.54720 1.00000 1.00000
```

```
ggplot(seda, aes(x = rural))+
geom_histogram(binwidth = 0.05)
```



The histogram is U shaped and shows an incredibly large number of the districts have 100% of students attending rural schools, there are also a very large number of school with 0% of students attending rural schools. The 3rd most common is 5% of students in rural schools, and the fewest number of districts have 95% of the students attending rural schools.

It seems natural that there would be many districts with 0% of students attending rural schools, as there are likely many districts that cover a smaller geographic area in urban settings; School districts in Brooklyn would likely not have any students attending rural schools. Opposingly, school districts that are in rural areas often cover a very large geographic area that is almost all rural, so it would be nearly impossible to have 95% of students attending urban schools, but it is quite possible that there is one rural school in a mostly urban district, so 5% of students attending that school makes plenty of sense. # Question 8

We will define a district as "rural" if 50% or more of the students in the district are in rural schools. Create a new variable that is equal to 0 if less than 50% of students in the district are in rural schools and equal to 1 if 50% or more of students are in rural schools. Call this variable "rural_ind". What percent of districts are rural based on this definition?

```
# code to create the rural_ind variable:
seda$rural_ind <- ifelse(seda$rural<0.5, 0, 1)

table(seda$rural_ind)
```

```
##
##    0    1
## 5861 6396
```

```
6396/12257
```

```
## [1] 0.5218243
```

52.1% of school districts in this dataset are rural.

Question 9

Use a hypothesis test to determine whether the difference in average math scores between rural and non-rural districts is statistically significant at the p<0.05 level. (HINT: you could use either the t.test() function in R or you can use a regression model). Briefly report your conclusion.

H0: average math scores are the same in rural and urban Districts HA: average math scores differ in rural and urban districts

```
rural<-seda%>%
        filter(rural_ind == 1)
urban<-seda%>%
        filter(rural_ind==0)

t.test(rural$avg_math, urban$avg_math)
```

```
##
##  Welch Two Sample t-test
##
## data:  rural$avg_math and urban$avg_math
## t = -10.005, df = 11267, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2414525 -0.1623420
## sample estimates:
## mean of x mean of y
##  5.478729  5.680626
```

Rural and Urban districts to not have the same average math scores. They differ significantly (p<0.05). Rural districts have a mean score of 5.48 and urban districts have a mean score os 5.68.

Question 10

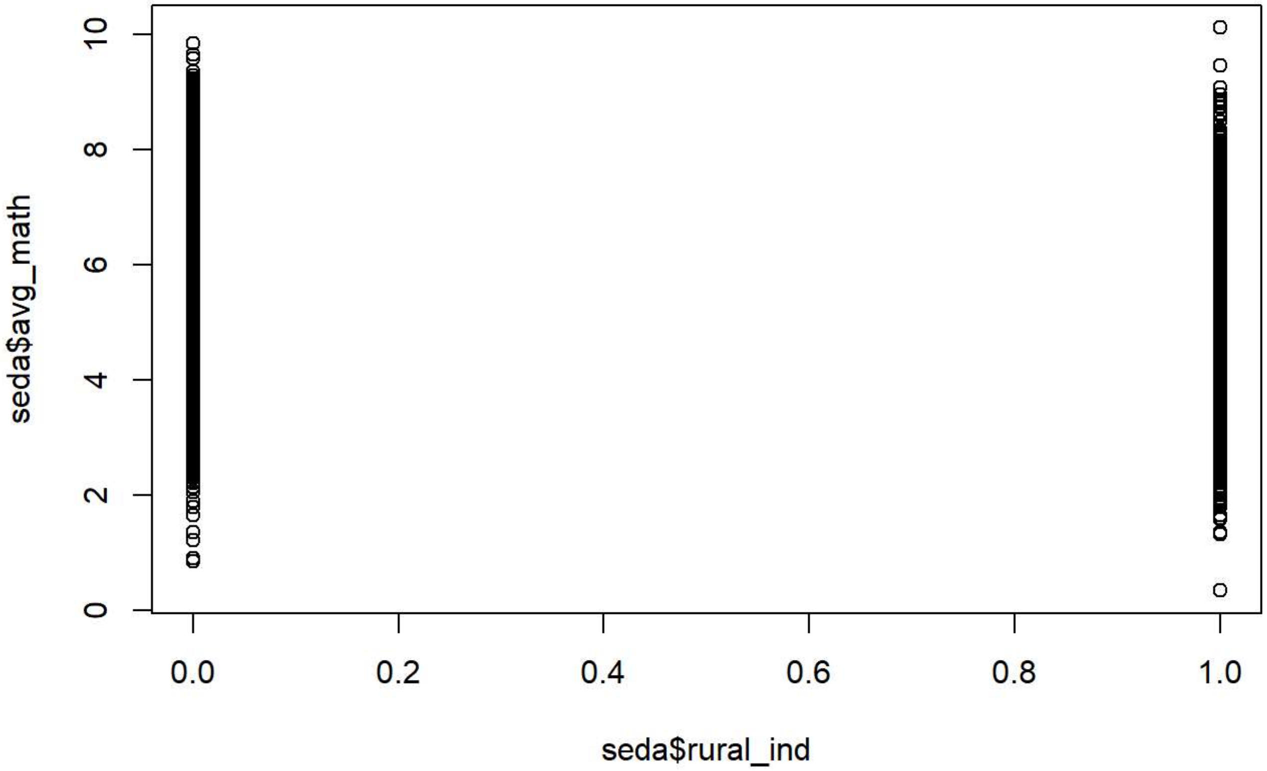
Now we'll compare the average math achievement in rural versus non-rural schools after controlling for differences in district SES. Estimate two regression models:

1. a simple regression model predicting average math scores from the rural_ind variable, and

2. a multiple regression model predicting average math scores as a function of the rural_ind variable and the avgsesall variables.

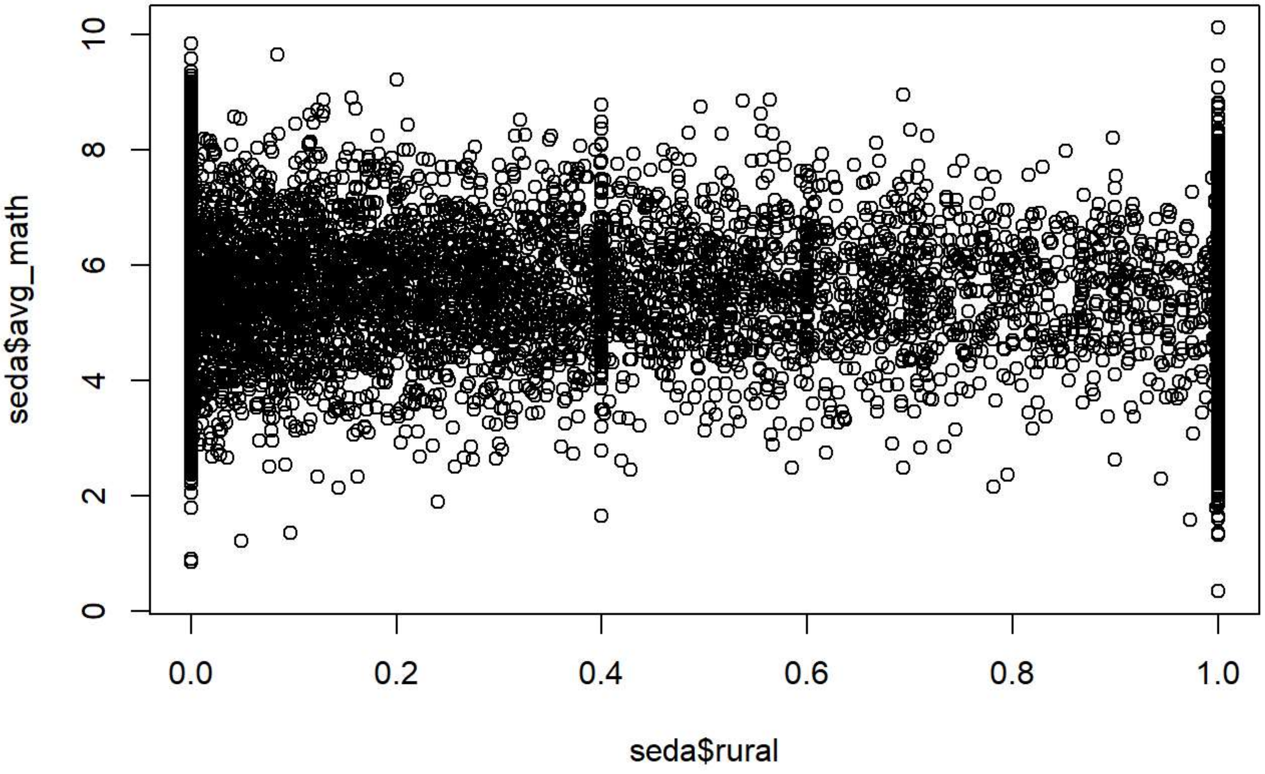
Display the two sets of model estimates side-by-side in a table (e.g., using the tab_model() function). Write 1 sentence interpreting each of the following estimates from model 2: intercept, coefficient for rural_ind, coefficient for avgsesall, and r-squared.

```
plot(seda$rural_ind, seda$avg_math)
```



```
# I cannot use a lm to plot a binary variable as there will not be a linear relationship. I will
use the variable rural instead
```

```
plot(seda$rural, seda$avg_math)
```



```
library(gridExtra) # for showing table side by side
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
rur_math_model1<-lm(avg_math~rural, data = seda)
rur_math_model2<-lm(avg_math~sesavgall+rural, data = seda)

tab_model(rur_math_model1, rur_math_model2,title = "Comparison of Simple and Multiple Linear Reg
ression Models",
          dv.labels = c("Simple Model (rural only)", "Multiple Model (SES and rural)"))
```

Comparison of Simple and Multiple Linear Regression Models

	Simple Model (rural only)			Multiple Model (SES and rural)		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	5.71	5.68 – 5.75	**<0.001**	5.35	5.33 – 5.37	**<0.001**

rural	-0.25	-0.30 – -0.21	**<0.001**	-0.18	-0.21 – -0.15	**<0.001**
sesavgall				0.96	0.95 – 0.98	**<0.001**
Observations	12257			12257		
R^2 / R^2 adjusted	0.010 / 0.010			0.550 / 0.550		

AI disclaimer I used github copilot to help understand how to generate two adjacent tables with tab_model()

Simple Model

- For districts with 0% of students in rural schools (100% urban), the predicted average math score is 5.71.
- For a 10% change in the proportion of rural students in the district, the average math score is predicted to change by negative 0.025 units.
- 1% of the variation in increase in math scores is explained by percent of students in rural schools in the district, likely because of the concentration in all-urban or all-rural schools.

Multiple Model

- When the school district has 0% rural students and the avg distrect SES score is 0, the predicted average math score of the district is 5.35.
- for each 10% increase in students attending rural schools in the district, the predicted avg math score would decrease by 0.018 holding constant the SES Scores. For each 1 unit increase in avg district SES score, the predicted avg math score would increase by 0.96, holding constant the percent of students in rural schools.
- 55% of the variation in increase in math scores is explained by the district's SES score and the % of students in rural schools.

Question 11

Is the difference in average math test scores between rural and non-rural districts statistically significant at the p<0.05 level after controlling for average district SES? How do you know?

Yes, the difference between rural and non-rural test scores is significantly different after controlling for average district SES. I know this because the the p-value for the rural variable in the second model is <0.001 meaning it is significant after controlling for SES.

Question 12

Now you'll examine whether the association between district SES and math scores differs among rural and non-rural districts. Estimate a multiple regression model predicting average math scores as a function of the rural_ind variable, the sesavgall variable, and the interaction between these two variables. Report the results of this model in a nicely formatted regression table.

```
rur_math_model3<-lm(avg_math~sesavgall+rural_ind + (sesavgall*rural_ind), data = seda)

tab_model(rur_math_model3, title = "Interactive model with SES and Rural")
```

Interactive model with SES and Rural

	avg math		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	5.31	5.29 – 5.33	**<0.001**
sesavgall	1.00	0.98 – 1.02	**<0.001**
rural ind	-0.11	-0.14 – -0.08	**<0.001**
sesavgall × rural ind	-0.10	-0.13 – -0.07	**<0.001**
Observations	12257		
R^2 / R^2 adjusted	0.551 / 0.550		

Question 13

Based on your regression model, does the association between average SES and average math scores differ in rural versus non-rural districts? How do you know? If the association is different, explain how the association differs. (HINT: it might help to create a graph!)

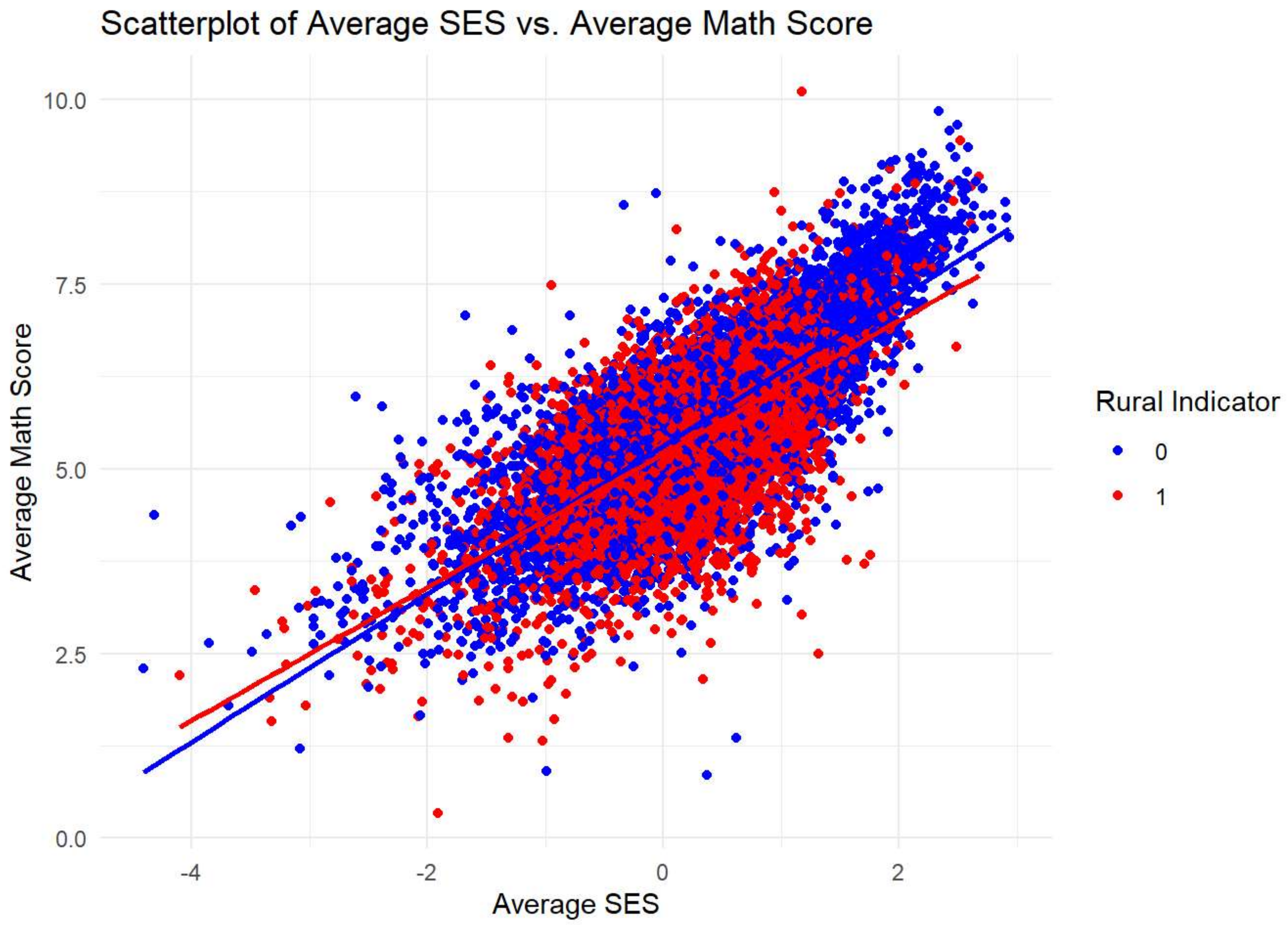
```
urban_lm <- lm(avg_math ~ sesavgall, data = subset(seda, rural_ind == 0))
rural_lm <- lm(avg_math ~ sesavgall, data = subset(seda, rural_ind == 1))
tab_model(urban_lm, rural_lm, title = "avg_math ~ Sesavgall compared in rural vs urban",
          dv.labels = c("Urban", "Rural"))
```

avg_math ~ Sesavgall compared in rural vs urban

	Urban			Rural		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	5.31	5.29 – 5.33	**<0.001**	5.20	5.18 – 5.22	**<0.001**
sesavgall	1.00	0.98 – 1.02	**<0.001**	0.90	0.88 – 0.93	**<0.001**
Observations	5861			6396		
R^2 / R^2 adjusted	0.626 / 0.626			0.436 / 0.436		

```
ggplot(seda, aes(x = sesavgall, y = avg_math, color = as.factor(rural_ind))) +
  geom_point() +
  geom_smooth(data = subset(seda, rural_ind == 0), aes(x = sesavgall, y = avg_math), method = "l
m", se = FALSE, color = "blue") +
  geom_smooth(data = subset(seda, rural_ind == 1), aes(x = sesavgall, y = avg_math), method = "l
m", se = FALSE, color = "red") +
  labs(title = "Scatterplot of Average SES vs. Average Math Score",
       x = "Average SES",
       y = "Average Math Score",
       color = "Rural Indicator") +
  scale_color_manual(values = c("0" = "blue", "1" = "red")) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



AI

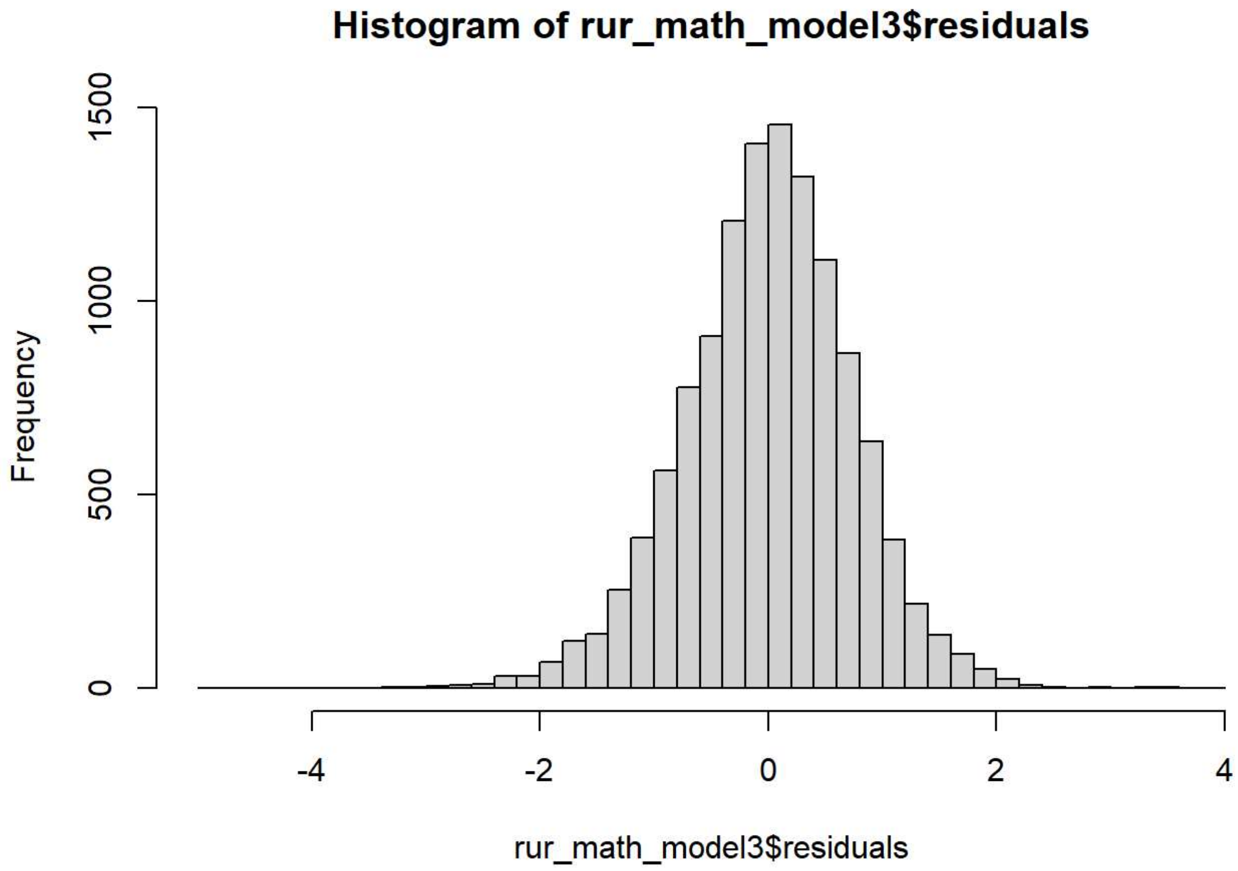
disclaimer I used copilot to help generate ggplot graph with two lines and different colored dots

Yes, the association between SES scores and avg math scores is different in rural and urban school districts. As seen in the plot above, the regression line for rural districts (red) is not as steep as the line for urban districts (blue). This implies that in urban districts, a change in SES predicts a greater change in avg math scores than in rural districts.

Question 14

Plot a histogram of the residuals from the model you estimated in Question #13. Briefly comment on whether the assumption that the residuals are normally distributed appears reasonable for these data.

```
hist(rur_math_model3$residuals, breaks = 40 )
```



Yes. The histogram of the residuals of from the interactive model appears to be very normal, so the assumption that residuals are normally distributed applies.